Importance-Weighted LLM Fine-Tuning for Relation Extraction

Tino Trangia, Teresa Lee, Raunak Sengupta
UCSD
{ttrangia, t1lee, r2sengupta}@ucsd.edu

Abstract

Relation Extraction (RE) detects and classifies semantic relationships between entities in text, but strong performance typically requires a sufficient quantity of high-quality labeled training data. In practice, training data often differs in distribution from real-world data, creating a distribution gap that limits model performance and makes increasing dataset size and quality costly. Weak supervision (including LLM-generated labels) offers scalability but introduces label noise and may exacerbate distribution mismatch. This project implements ATLANTIS, an importance-weighted weak-to-strong learning method for supervised fine-tuning, and applies it to sentence-level relation extraction. Across encoder–decoder (Flan-T5) and decoder-only (Qwen2) models on SemEval-2010 Task 8 and CoNLL2004, importance weighting yields small and setting-sensitive changes relative to a standard supervised fine-tuning baseline, while remaining compatible with both model families. Code: https://github.com/ttrangia/DSC253_Importance_Weighted_Relation_Extraction

1 Introduction

Relation Extraction (RE) is the task of detecting and classifying semantic relationships between entities in text. In the sentence-level setting, the input consists of a single sentence containing two marked entity mentions, and the model predicts a relation label describing the semantic interaction between them. Sentence-level RE serves as a core component of larger information extraction systems, enabling structured knowledge construction and downstream reasoning applications.

High-performing RE systems typically require sufficiently large quantities of high-quality labeled training data. However, expert annotation is expensive and scarce, especially in specialized domains such as medicine or law. Even when labeled datasets are available, their distribution may differ from real-world data, creating a distribution gap that limits model generalization. Simply increasing dataset size is often costly and does not necessarily resolve this mismatch.

Weak supervision offers a practical path toward scalability. Large language models (LLMs) can generate substantial volumes of pseudo-labeled data at low cost, enabling rapid dataset expansion. However, weak labels are frequently noisy and may exhibit even larger distributional discrepancies than manually curated data. Naively fine-tuning on noisy labels can degrade performance rather than improve it, raising a central question: how can we leverage the scale of weak supervision while mitigating its adverse effects?

ATLANTIS (Liu et al., 2025) proposes a weak-to-strong learning framework that addresses distribution mismatch in supervised fine-tuning through importance weighting. Instead of treating all training examples equally, ATLANTIS adjusts the optimization direction using per-example influence weights computed from the probability gap between a small base model, a reference model, and a large base model. Originally evaluated in instruction-tuning settings across broad knowledge and preference benchmarks, ATLANTIS demonstrates that distribution-aware reweighting can improve model alignment and performance under dataset mismatch.

In this work, we adapt ATLANTIS to sentence-level relation extraction and investigate whether importance-weighted fine-tuning can improve robustness under weak supervision. Specifically, we examine both fully supervised settings and weakly labeled scenarios in which controlled noise is introduced into training labels. Our central hypothesis is that importance weighting can reduce the negative impact of noisy supervision by biasing training toward examples that better align with a reference-informed distribution, thereby narrowing

the effective distribution gap.

Our contributions are as follows:

- We implement ATLANTIS-style importance weighting from scratch and adapt it to sentence-level relation extraction for both encoder–decoder and decoder-only architectures.
- We empirically evaluate the method on SemEval-2010 Task 8 and CoNLL2004 under both gold-labeled and weakly labeled training settings.
- We analyze the extent to which importance-weighted fine-tuning mitigates label noise relative to a standard supervised fine-tuning baseline.

2 Background and Motivation

2.1 Supervised Fine-Tuning and the Optimal Distribution

Supervised fine-tuning (SFT) aims to align a model's output distribution with a target data distribution. In the formulation presented by ATLANTIS (Liu et al., 2025), the ideal objective is to minimize the divergence between the model distribution p_θ and an optimal real-world distribution p^*:

$$\mathcal{L}(\theta) = -\mathbb{E}_{x\sim q(\cdot),\, y\sim p^*(\cdot|x)} \log p_\theta(y|x). \quad (1)$$

Here, p^* denotes an idealized distribution that fully reflects real-world language patterns. In practice, training data are drawn from a finite empirical distribution p_d, which may differ from p^*. ATLANTIS argues that this mismatch constrains model capacity and generalization when optimizing directly under p_d (Liu et al., 2025).

2.2 Importance Sampling Perspective

Because p^* is not directly accessible and cannot be sampled or evaluated explicitly, ATLANTIS reframes the objective using importance sampling. Conceptually, importance sampling allows expectations under a target distribution to be estimated using samples from another distribution by applying a corrective ratio.

Rewriting the ideal objective under an accessible reference distribution p_r yields:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y)\sim p_r} \frac{p^*(y|x)}{p_r(y|x)} \log p_\theta(y|x). \quad (2)$$

The central difficulty is that the ratio $\frac{p^*(y|x)}{p_r(y|x)}$ is incalculable, since p^* itself is unknown (Liu et al., 2025). Therefore, ATLANTIS focuses on estimating this ratio indirectly using model distributions.

2.3 Weak-to-Strong Proportionality Assumption

ATLANTIS introduces a weak-to-strong proportionality assumption relating a large base model p_b^L, a small base model p_b^S, and a reference distribution p_r. The key assumption is:

$$\frac{p^*(y|x)}{p_b^L(y|x)} \propto \frac{p_r(y|x)}{p_b^S(y|x)}. \quad (3)$$

This implies:

$$\frac{p^*(y|x)}{p_r(y|x)} \propto \frac{p_b^L(y|x)}{p_b^S(y|x)}. \quad (4)$$

Under this proportionality, the intractable importance ratio can be approximated using probabilities from the large and small models (Liu et al., 2025). The resulting framework adjusts optimization direction during fine-tuning by assigning higher influence to examples whose probability shifts are more informative.

2.4 Relevance to Weak Supervision in Relation Extraction

In sentence-level relation extraction, weak supervision can be introduced through controlled label noise or automatically generated labels. In such cases, the empirical training distribution deviates further from the ideal distribution due to incorrect or noisy labels. This setting provides a natural testbed for evaluating whether distribution-aware weighting mechanisms can mitigate the impact of noise.

Our work applies the ATLANTIS framework to sentence-level relation classification to examine whether importance-weighted optimization reduces degradation under noisy supervision while maintaining compatibility with standard fine-tuning procedures.

3 Related Work

Our work connects sentence-level relation extraction, weak supervision for information extraction, and weak-to-strong learning methods for fine-tuning large language models.

3.1 Sentence-Level Relation Extraction

Sentence-level relation extraction (RE) is typically formulated as multi-class classification over predefined relation types given a sentence and two marked entities. SemEval-2010 Task 8 (Hendrickx et al., 2010) is a widely used benchmark defining 19 mutually exclusive semantic relations, and remains a standard testbed for evaluating relation classification systems. CoNLL2004 (Roth and Yih, 2004) is another benchmark frequently used in joint entity and relation extraction research and can be adapted to sentence-level relation classification when entity spans are provided. These datasets offer controlled environments for studying learning behavior under different supervision regimes.

Beyond sentence-level settings, document-level benchmarks such as DocRED (Yao et al., 2019) introduce cross-sentence reasoning challenges. Although our work focuses on sentence-level RE, these benchmarks motivate examining training strategies that remain robust under distributional variation.

3.2 Weak Supervision for Relation Extraction

Weak supervision has long been used to scale relation extraction beyond manually labeled datasets. Distant supervision (Mintz et al., 2009) aligns text with knowledge bases to heuristically generate relation labels, but introduces systematic noise due to imperfect alignment assumptions. More general weak supervision frameworks such as Snorkel (Ratner et al., 2017) treat labeling functions as noisy sources and learn to aggregate them into probabilistic training labels. These approaches highlight the central trade-off between scale and label quality.

In recent years, large language models have been used to generate pseudo-labels for downstream tasks, including relation extraction, offering a flexible alternative to rule-based distant supervision. However, noisy labels from LLMs can degrade supervised fine-tuning when treated uniformly, motivating training procedures that explicitly account for example reliability.

3.3 LLMs and Instruction-Based Relation Extraction

Large language models enable generative and instruction-based formulations of relation extraction. Instead of training discriminative classifiers over fixed label indices, models can be prompted to produce relation labels directly as text. Instruction-tuning paradigms (Wei et al., 2022; Ouyang et al., 2022) have demonstrated strong generalization across tasks, and similar prompting strategies have been explored for information extraction. This generative framing allows unified treatment across encoder–decoder and decoder-only architectures and aligns with our implementation design. Recent work has studied LLM performance on information extraction specifically, demonstrating the competitiveness of few-shot and supervised finetuning methods for such tasks (Wadhwa et al., 2023).

3.4 Weak-to-Strong Learning and Data Selection

Weak-to-strong learning investigates whether stronger models can benefit from weaker supervision signals. Burns et al. (Burns et al., 2023) study weak-to-strong generalization and show that strong models can sometimes learn from weaker supervisory signals, but naive training may not fully exploit this potential.

A related line of work focuses on data selection and filtering for instruction tuning. LESS (Xia et al., 2024) estimates training data influence through optimizer-aware signals to select subsets of examples for targeted fine-tuning. Superfiltering (Li et al., 2024) explores weak-to-strong data filtering strategies for improving instruction tuning efficiency. These approaches emphasize that not all training examples contribute equally to model improvement.

ATLANTIS (Liu et al., 2025) differs from data selection methods by reweighting all training examples instead of discarding subsets. It introduces an importance sampling framework that estimates per-example influence weights using probability gaps between small and large models. While originally evaluated in instruction-tuning contexts, its formulation is general and can be applied to structured prediction tasks. Our work extends ATLANTIS to sentence-level relation extraction under weak supervision, evaluating whether importance-weighted optimization improves robustness relative to standard supervised fine-tuning.

4 Methodology

We adapt ATLANTIS-style importance-weighted supervised fine-tuning (Liu et al., 2025) to sentence-level relation extraction (RE). This section describes the task formulation, baseline training objective, the computation of ATLANTIS influence

weights, and the weak supervision setup used in our experiments.

4.1 Task Formulation

We formulate sentence-level RE as multi-class classification. Each instance consists of a sentence x containing two marked entity mentions and a relation label $y \in \mathcal{R}$.

For SemEval-2010 Task 8, the original label space contains 19 relation classes including *Other*. In our experiments, we optionally collapse directional variants into 10 direction-agnostic relation types to reduce directional ambiguity. For CoNLL2004, the label space contains 5 relation types.

We treat RE as an instruction-style generation task. Given a sentence with marked entities, the model receives a prompt of the form:

> *"Classify the semantic relation between the marked entities in this sentence: [sentence]"*

The model generates the relation label as text. Evaluation is performed using Macro-F1.

4.2 Baseline Supervised Fine-Tuning

The baseline objective follows standard supervised fine-tuning (SFT). Given training distribution p_d, we minimize:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim p_d} \log p_\theta(y \mid x), \qquad (5)$$

where p_θ denotes the model distribution over relation labels.

We evaluate both encoder–decoder models (Flan-T5) and decoder-only models (Qwen2). In both cases, full fine-tuning is performed using the hyperparameters specified in our experimental setup.

4.3 ATLANTIS Importance Weighting

ATLANTIS introduces per-example influence weights computed from three model distributions:

- p_b^L: large base model (to be fine-tuned),
- p_b^S: small base model,
- p_r: reference model.

Reference Model Instantiation. In our implementation, the reference distribution p_r is instantiated as a small model fine-tuned on the gold training split of the target relation extraction dataset. We also evaluated using the corresponding instruction-tuned checkpoint as the reference model. Empirically, both configurations produced nearly identical influence weight distributions and downstream performance. For consistency and reproducibility, we report results using the instruction-tuned model as p_r.

Following Algorithm 1 in ATLANTIS (Liu et al., 2025), the weight for example (x_i, y_i) is:

$$W_i = \frac{p_b^L(y_i \mid x_i)}{p_b^S(y_i \mid x_i)} \cdot p_r(y_i \mid x_i). \qquad (6)$$

Weights are precomputed and cached prior to fine-tuning. The large model is then trained using a weighted cross-entropy objective:

$$\mathcal{L}_{\text{ATL}}(\theta) = -\sum_{i\in\mathcal{B}} W_i \log p_\theta(y_i \mid x_i), \qquad (7)$$

where $\mathcal{B}$ denotes a minibatch.

Probability Computation. Since relation labels are produced as text, we compute example probabilities using the conditional log-likelihood of the correct label string given the input prompt. For decoder-only models, this is obtained by summing the token-level log-probabilities of the label tokens. For encoder–decoder models, probabilities are computed under teacher forcing over the target sequence. In all cases, we use the total log-likelihood of the label sequence without length normalization.

Weight Normalization. To ensure stable optimization, importance weights are normalized prior to training so that their mean equals 1.0 across the training set. When enabled, weight clipping is applied before normalization to prevent extreme values from dominating the loss.

4.4 Weak Supervision Setup

To simulate weak supervision, we introduce controlled label noise into the SemEval training set by replacing gold labels with incorrect labels at rates between 20% and 50%. Let p_d^{noise} denote the resulting corrupted training distribution.

We compare:

- Standard SFT on p_d^{noise}.
- ATLANTIS-weighted SFT on p_d^{noise}.

This setup allows us to evaluate whether importance weighting mitigates performance degradation caused by noisy supervision.

4.5 Implementation Scope

Our implementation follows the three-model structure required by ATLANTIS and applies importance weighting to sentence-level relation classification only. We do not address document-level relation extraction or end-to-end entity detection.

5 Experimental Setup

5.1 Datasets

We evaluate relation extraction performance on:

- **SemEval-2010 Task 8 (gold labels):** Sentence-level classification of mutually exclusive semantic relations with 8K training and 2.7K test samples. Entity pairs are tagged in the input. The dataset contains 19 relation classes (including "Other"), direction-collapsed to 10.
- **SemEval-2010 Task 8 (weakly labeled):** A weakly labeled training set used to simulate real-world use of weak supervision, with adjustable noise rates from 20–50% through random correction. Labels were generated using Qwen2-1.5B.
- **CoNLL2004:** A joint NER+RE dataset with 922 training and 288 test samples and 5 relation classes. The data includes entity positions and relation type/positions. We parse the structured format into natural sentences and split samples with multiple entity/relation pairs into separate training instances.

5.2 Models

We evaluate:

- **Flan-T5** (encoder–decoder): small (77M) and base (220M).
- **Qwen2** (decoder-only): 0.5B and 1.5B; and 1.5B and 7B configurations. These match the models used in the original ATLANTIS paper.

5.3 Training Details

Flan-T5. We use full fine-tuning with 8 epochs, learning rate 3×10^{-4}, and batch size 32.

Qwen2. We use full fine-tuning with 3 epochs and learning rate 2×10^{-5}. For the 0.5B/1.5B configuration, we use batch size 8 with gradient accumulation 2. For the 1.5B/7B configuration, we use batch size 1 with gradient accumulation 16.

Model	Setting	SFT	ATLANTIS
Flan-T5	SemEval (gold)	0.849	0.854
Qwen2-1.5B	SemEval (gold)	0.850	0.853
Qwen2-7B	SemEval (gold)	0.885	0.877

Table 1: Macro-F1 on SemEval-2010 Task 8 (gold labels).

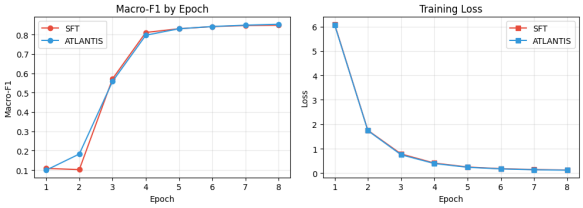

Figure 1: Learning curves for Flan-T5 on SemEval (gold labels)

Evaluation Metric. We report Macro-F1 as shown in the project results for each test set.

6 Results

6.1 SemEval-2010 Task 8 (Gold)

Table 1 summarizes Macro-F1 results on gold-labeled SemEval. Figure 1 shows the Macro-F1 and training loss over each epoch for Flan-T5 in this setting.

6.2 CoNLL2004

Table 2 summarizes Macro-F1 results on CoNLL2004. Figure 2 and Figure 3 show the learning curves for Qwen2-1.5B and Qwen2-7B respectively in this setting.

6.3 SemEval-2010 Task 8 (Weakly Labeled; 20% Noise)

We also evaluate Qwen2-1.5B on a weakly labeled SemEval training set with 20% noise. Table 3 reports Macro-F1.

6.4 Takeaways

Across the evaluated settings, ATLANTIS-style importance weighting is compatible with both

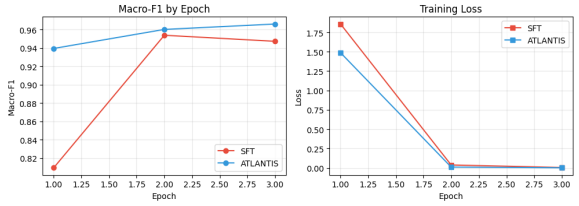

Figure 2: Learning curves for Qwen2-1.5B on CoNLL2004

Model	SFT	ATLANTIS
Qwen2-1.5B	0.947	0.966
Qwen2-7B	0.982	0.982

Table 2: Macro-F1 on CoNLL2004.

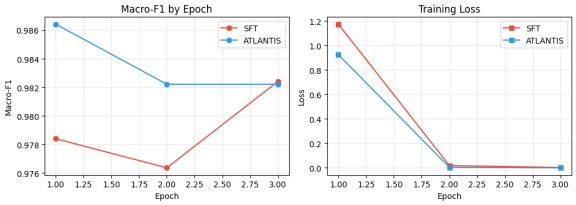

Figure 3: Learning curves for Qwen2-7B on CoNLL2004

encoder–decoder and decoder-only fine-tuning for relation extraction. The effect on finetuned performance appears to be small and sensitive to experimental settings, such as epochs trained, learning rate, and noise rate. From inspecting the learning curves, we see similar patterns across epochs, with only small deviations in Macro-F1.

Overall, the results suggest that while importance weighting may be useful for improving fine-tuning on noisy datasets, it does not achieve the same improvements as the original paper. We discuss potential causes of this in the limitations section. Standard supervised fine-tuning remains competitive where clean data is available, however methods that can improve zero/few-shot relation extraction methods remain desirable.

7 Limitations

Setting mismatch. ATLANTIS was originally designed for general instruction tuning with official instruction-tuned models as reference and evaluation on mixed benchmarks, whereas we apply it to a narrower task with a small fine-tuning dataset. This may explain why the performance gains were not realized, as the original tasks had more variance and potential for improvement than our relation extraction task.

Pipeline scope. Datasets are evaluated on annotated entity pairs only, bypassing the entity detection component of full RE pipelines and inflating absolute F1 relative to expected end-to-end settings.

Lack of SOTA comparisons. We compare relative improvements against an SFT baseline but do not measure absolute state-of-the-art performance; the approach could be combined with other methods (e.g., data selection) for a more realistic end-to-end model.

Setting	SFT (uniform)	ATLANTIS
SemEval weakly labeled (20% noise)	0.809	0.824

Table 3: Macro-F1 on SemEval-2010 Task 8 with weak labels (20% noise).

8 Future Work

Our findings open several promising directions for future investigation. First, broader empirical evaluation across diverse benchmarks including ADE, TACRED, and NYT, as well as cross-domain settings, would test whether importance weighting generalizes beyond the datasets we examined.

Second, systematic ablation studies varying noise rates, model size pairs, weight clipping thresholds, and reference model choices would clarify the conditions under which weighting provides measurable benefit.

Third, extending the framework to document level relation extraction on Re-DocRED would test its applicability to cross sentence reasoning tasks, while multi task information extraction would require developing joint weighting strategies for models performing both entity and relation extraction.

Finally, moving from controlled noise to realistically weak supervision such as LLM generated labels with systematic errors or distantly supervised data would provide stronger evidence for the method's utility in practical settings where training data is inherently noisy and imperfect.

Acknowledgments

This report is written as part of DSC 253.

References

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs

of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *arXiv preprint arXiv:2402.00530*.

Yi Liu, Guoyin Wang, Shicheng Li, Feifan Song, and Xu Sun. 2025. Atlantis: Weak-to-strong learning via importance sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Alexander Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment*, volume 11, pages 269–282.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL-2004)*. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

A Appendix: Prompts and Data Formatting

We use an instruction-style prompt for relation classification:

> *"Classify the semantic relation between the entities [entity 1] and [entity 2] in this sentence: . . . "*

For CoNLL2004, we parse structured dictionaries into natural sentences and split samples with multiple entity/relation pairs into separate instances.